IEEE Conference on

PATTERN RECOGNITION AND IMAGE PROCESSING

AUGUST 3-5, 1981

PROCEEDINGS

proceedings

PRIP 81

AUGUST 3-5, 1981
DALLAS, TEXAS

IEEE
COMPUTER
SOCIETY
PRESS

IEEE Catalog No. 81CH1595-8
Library of Congress No. 80-83214
Computer Society No. 352

Order from: IEEE Computer Society
Post Office Box 80452
Worldway Postal Center
Los Angeles, CA 90080

IEEE Service Center
445 Hoes Lane
Piscataway, NJ 08854

The Institute of Electrical and Electronics Engineers, Inc.

Conference Chairman's Message

It is with a great deal of pride and pleasure that I present this volume to the pattern recognition and image processing community. The size of the volume is indicative of the substantial research in progress, and the excellence of the presentations will attest to the high quality of the research being pursued. The broad participation from academic, industrial, and government personnel in the conference clearly shows the significance and utility of our research.

I would like to take this opportunity to thank my Conference Committee and the Program Committee for all the support they have provided. Everyone is grateful to Professor Azriel Rosenfeld for having arranged an excellent program. Also, I would like to convey my sincere thanks to Professors D. Green and N. Badler, and Dr. T. Lucido for their cooperation in arranging this conference in conjunction with SIGGRAPH '81. Again, it is a pleasure to thank the IEEE Computer Society for sponsoring this conference and the IEEE Computer Society staff for their marvelous help in making various arrangements.

J. K. Aggarwal
General Chairman

Conference Committee

Chairman

J.K. Aggarwal
The University of Texas
Austin, Texas

Finance and Vice-Chairman

Larry S. Davis
The University of Texas
Austin, Texas

Publication and Publicity

Kenneth R. Diller
The University of Texas
Austin, Texas

Exhibits

Marg Knox
The University of Texas
Austin, Texas

Program Chairman

Azriel Rosenfeld
University of Maryland
College Park, Maryland

Local Arrangements

Stephen A. Underwood
The University of Texas
Arlington, Texas

Program Committee

Chairman

Azriel Rosenfeld
University of Maryland
College Park, Maryland

Committee

Norman Badler
University of Pennsylvania
Philadelphia, Pennsylvania

Ruzena Bajcsy
University of Pennsylvania
Philadelphia, Pennsylvania

Richard O. Duda
Fairchild Central Research & Development
Palo Alto, California

Herbert Freeman
Rensselaer Polytechnic Institute
Troy, New York

K.S. Fu
Purdue University
Lafayette, Indiana

Robert M. Haralick
Virginia Polytechnic Institute
& State University
Blacksburg, Virginia

Thomas S. Huang
University of Illinois
Urbana, Illinois

Martin Levine
McGill University
Montreal, Canada

H.H. Nagel
Universität Hamburg
West Germany

David Nitzan
SRI International
Menlo Park, California

Theodosios Pavlidis
Bell Laboratories
Murray Hill, New Jersey

Judith S. Prewitt
National Institutes of Health
Bethesda, Maryland

William B. Thompson
University of Minnesota
Minneapolis, Minnesota

Conference on Pattern Recognition and Image Processing

Dallas, Texas
August 1-5, 1981

Table of Contents

原书模糊

Session A: Classification

L.N. Kanal
University of Maryland

A COMMITTEE MACHINE WITH LOWER COMMITTEES

Ryuzo Takiyama

Department of Visual Communication Design
Kyushu Institute of Design
Minami-ku, Fukuoka, 815 Japan

ABSTRACT A committee machine with several lower committees, which is called a two-level committee machine, is proposed to improve the pattern classification power of the usual committee machine, and the learning algorithm for it is described. The discriminant function realized by the two-level committee machine can be considered as the general piecewise linear discriminant function which includes Chang's definition[12]. The proposed algorithm is a kind of error-correction procedure, and the learning procedures of the usual committee machine and the perceptron are clearly explained as special cases of the proposed algorithm.

I. INTRODUCTION

It is convenient to use the concept of the discriminant function(DF), f, to state the 2-class pattern classification problem in pattern recognition. If f is piecewise linear, it is called the piecewise linear discriminant function(PLDF). Although the usefulness of the PLDF has already been pointed out, and there has been much research on synthesizing it, satisfactory results have not yet been obtained, and the problem has rarely been treated in its general form.

Most research done so far relates to the nearest neighbor classifiers[1][2], the committee machine[3][4][5][6], or other types of networks of linear classifiers[7][8][9][10][11]. The DF's derived from those mentioned above are restricted or special PLDF's. Although Chang[12] has treated the PLDF in more general fashion, the analytical representation of the PLDF has not been presented and the synthesis procedure for it is rather heuristic. It is interesting from not only a practical but a theoretical point of view to investigate the PLDF in its general form.

In this paper, we are motivated by the committee machine, and aim to describe the PLDF in more general form. Thus, we propose a two-level committee machine(TLCM)[13]. The TLCM is a committee machine whose committee members are also committee mahcines. The TLCM defines a wider family of PLDF's which includes Chang's definition as a special case.

We present an analytical expression of the DF realized by the TLCM, and give a learning procedure for the TLCM. As results, we present an analytical expression of the general PLDF and a learning method to find a general PLDF from sample patterns.

II. THE TWO-LEVEL COMMITTEE MACHINE

2.1 The committee machine

In what follows patterns are divided into two classes, C_1 and C_2, and are n-tuples, i.e.

$$x = (x_1, \ldots, x_n). \tag{1}$$

For notational simplicity the last or n-th component of each pattern is the constant 1.

The perceptron is a two-class classifier denoted by L(x) and defined by

$$L(x) = \begin{cases} 1 & \text{if } w \cdot x > 0 \\ -1 & \text{if } w \cdot x \leq 0, \end{cases} \tag{2}$$

where w, called a weight vector, is in R^n, "$\cdot$" means the inner product, and $w \cdot x$ is the DF on which the perceptron makes the decision(classification). The components of x are called the inputs to the perceptron, and L(x) is called the response of the perceptron.

A more powerful classifier than a perceptron is obtained by combining several perceptrons called committee members and a vote-taking perceptron. The components of the pattern vector are the inputs to each committee member, and the responses of the committee members together with the constant one are the inputs to the vote-taking perceptron. The response of a committee is the response of its vote-taking perceptron, and the weight vector of the vote-taking perceptron determines what is called the logic of the committee.

2.2 The TLCM

A TLCM proposed in this paper is a committee machine in which each committee member is also a commiittee machine. That is, a TLCM consists of several committee machines as committee members and a vote-taking perceptron. Committee machines which constitute committee members are called lower committees whereas the committee machine which makes the final decision is called the upper committee. Inputs to the vote-taking perceptron of the upper committee are responses of lower committees(outputs of vote-taking perceptrons in lower committees).

In order to characterize a TLCM, several parameters are to be determined. These parameters are classified into two kinds,i.e. outer parameters and inner parameters. The number of lower committee members, the number of committee members of each lower committee, and logics of lower and upper committees are outer parameters. Weight vectors of committee members of each lower committee are inner parameters.

Outer parameters determine the form of the

DF realized whereas inner parameters determine the function of the machine as far as the functional form given by outer parameters. In most classifier design problems, it is assumed that outer parameters are known in advance, and the problem is to find inner parameters in such a way that the function of the machine works well.

III. THE DISCRIMINANT FUNCTION OF THE TLCM

Let us consider a TLCM with J committee members. As was mentioned in the previous section, each member is a committee machine with respective logic. In these J committees, let the j-th committee be denoted by C(j), j=1,...,J, and let C(j) have K_j lower committee members.

Further let the weight vector of the k-th committee member of C(j) be

$$w_k^j = (w_{k1}^j,\dots,w_{kn}^j),\ j=1,\dots,K_j, \tag{3}$$

and let v^j be the weight vector of the vote-taking perceptron of C(j). v^j has K_j+1 components, i.e.

$$v^j = (v_1^j,\dots,v_{K_j+1}^j), \tag{4}$$

the first K_j correspond to the responses from lower committee members in C(j) and the last to the input constant one.

Defining the threshold mapping $\Theta[\cdot]$ by

$$\Theta[u] = \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{if } u \leqq 0, \end{cases} \tag{5}$$

then L(x) defined by (2) can be written

$$L(x) = \Theta[w\cdot x]. \tag{6}$$

Therefore the weighted sum of inputs to the vote-taking perceptron in C(j) can be represented as

$$g_1^j(x) = \sum_{k=1}^{K_j} v_k^j \Theta[w_k^j\cdot x] + v_{K_j+1}^j, \tag{7}$$

and the response of C(j) is 1 if $g_1^j(x) > 0$, and -1 if $g_1^j(x) \leqq 0$. That is, $g_1^j(x)$ given by (7) is the DF realized by C(j), which is a kind of PLDF.

Inputs to the vote-taking perceptron of the upper committee are responses of lower committees, $\Theta[g_1^j(x)]$, j=1,...,J. Let the logic of the upper committee be

$$u = (u_1,\dots,u_J,u_{J+1}), \tag{8}$$

where the first J correspond to the responses from committee members and the last to the input one. Then the weighted sum of inputs to the upper vote-taking perceptron can be represented by

$$g_1(x) = \sum_{j=1}^{J} u_j \Theta[g_1^j(x)] + u_{J+1}, \tag{9}$$

which is the DF realized by the TLCM.

It is shown in [13] that the DF given by (9) can be considered as the general PLDF which includes Chang's definition.

As has been mentioned in Section II, parameters J, K_j, v^j and u in $g_1(x)$ are outer parameters, and vectors w_k^j are inner parameters which are to be found by learning or other methods.

IV. AN EASILY LEARNABLE FORM OF THE DF OF THE TLCM

Let us assume that the outer parameters are given and fixed. Under this circumstance we determine inner parameters by learning. Although the representation of the DF of the TLCM, $g_1(x)$, given by (9), is easy to understand, it is not necessarily desirable to introduce the learning algorithm for obtaining inner parameters. Therefore we represent it in another form.

We start with representing inner parameters in compact form. We define following vectors with nK_j-dimensions.

$$P_j = (w_1^j,\dots,w_{K_j}^j),\ j=1,\dots,J \tag{10}$$

$$X_j = (\underbrace{x,\dots,x}_{K_j}), \tag{11}$$

where P_j is the parameter vector and X_j is the augmented pattern vector with respect to the j-th lower committee member C(j). We introduce the following functions $t_k(X_j,P_j)$, $j=1,\dots,K_j$ and a diagonal matrix $T(X_j,P_j)$ for each j.

$$t_k(X_j,P_j) = \frac{1}{2} + \frac{1}{2}\Theta[w_k^j\cdot x]\Theta[g_1^j(X_j,P_j)],\ k=1,\dots,K_j \tag{12}$$

$$T(X_j,P_j) = \begin{pmatrix} t_1(X_j,P_j)E & & 0 \\ & \ddots & \\ 0 & & t_{K_j}(X_{K_j},P_{K_j})E \end{pmatrix}, \tag{13}$$

where $g_1^j(X_j,P_j)$ in (12) is the expression defined by (7) but denoted in a manner which explicitly represents the parameters, and E in (13) denotes the unit matrix of order n. Hence $T(X_j,P_j)$ is a diagonal matrix of order nK_j.

If we define a function

$$P_jT(X_j,P_j)X_j^t, \tag{14}$$

where t denotes the transposition, then we see that the function defined by (14) is the DF realized by C(j)[6].

In order to introduce a learnable representation of the DF of the TLCM, i.e. the general PLDF, we define the following $n(K_1+K_2+\dots+K_J)$ dimensional vectors.

$$P = (P_1,\dots,P_J) \tag{15}$$

$$X = (X_1,\dots,X_J), \tag{16}$$

where P is the parameter vector and X is the augmented pattern vector with respect to the overall committee machine. Further we define functions

$$s^j(X,P) = \frac{1}{2} + \frac{1}{2}\Theta[P_jT(X_j,P_j)X_j^t]\Theta[g_1(X,P)] \quad j=1,\dots,J \tag{17}$$

and a diagonal matrix of order $n(K_1+\dots+K_J)$,

$$S(X,P) = \begin{pmatrix} s^1(X,P)T(X_1,P_1) & & 0 \\ & \ddots & \\ 0 & & s^J(X,P)T(X_J,P_J) \end{pmatrix}. \tag{18}$$

$g_1(X,P)$ in (17) is the expression defined in (9) but is denoted in a manner which explicitly represents the parameters, and $T(X_j,P_j)$ in (18) is given by (13).

By using (15)-(18), we define the following function:

$$g(X,P) = PS(X,P)X^t. \tag{19}$$

Then we see that g(X,P) is the DF realized by the TLCM, i.e. the general PLDF[13]. Henceforth we will discuss the learning algorithm based on g(X,P).

Before going to the next section, some remarks will be given. Since $\Theta[cu] = \Theta[u]$ for $c > 0$, it holds that

$$s^j(X,cP) = s^j(X,P) \quad (20)$$

and

$$g(X,cP) = cg(X,P), \quad (21)$$

that is, $s^j(X,P)$ is a zero order homogeneous function and $g(X,P)$ is a linear homogeneous function with respect to P. Further defining

$$\tilde{S}(X,P) = \begin{cases} S(X,P) \\ -S(X,P) \end{cases}$$

and

$$\tilde{g}(X,P) = P\tilde{S}(X,P)X^t, \quad (22)$$

then if C_1 and C_2 can be correctly classified by the TLCM, there exists a parameter vector which satisfies

$$\tilde{g}(X,P) > 0, \quad X \in C_1 U C_2. \quad (23)$$

V. A LEARNING ALGORITHM FOR THE TLCM

A learning algorithm of the TLCM is performed by feeding patterns in sequence as is the case for the perceptron. Let an infinite sequence of learning patterns be A, i.e.

$$A = X(1), X(2), \ldots, X(r), \ldots . \quad (24)$$

The sequence A satisfies the same conditions as those of the perceptron.

Let us consider an algorithm based on a gradient method that seeks a minimum of a certain criterion function. The criterion function R(X,P) is chosen in such a way that it is minimum when $\tilde{g}(X,P) > 0$. The gradient descent procedure is then

$$P(r+1) = P(r) - a_r \nabla_P R|_{P(r),X(r)}, \quad (25)$$

where $\nabla_P R$ is the gradient of R with respect to P, P(r) is the value of P at the r-th iteration and a_r is a predetermined positive constant.

Let

$$R = \frac{1}{2}\{|\tilde{g}(X,P)| - \tilde{g}(X,P)\} \quad (26)$$

as is the case for the perceptron, then (25) can be rewritten as follows.

$$P(r+1) = P(r) + \begin{cases} 0, & \tilde{g}(r) > 0 \\ a_r \nabla_P \tilde{g}|_{P(r),X(r)}, & \tilde{g}(r) \leqq 0 \end{cases} \quad (27)$$

where $\nabla_P \tilde{g}$ is the gradient of $\tilde{g}(X,P)$ and $\tilde{g}(r)$ stands for $\tilde{g}(X(r),P(r))$. $\nabla_P \tilde{g}$ becomes,

$$\nabla_P \tilde{g} = \nabla_P (P\tilde{S}(X,P)X^t)$$
$$= (\tilde{S}(X,P)X^t)^t + P\nabla_P(\tilde{S}(X,P)X^t). \quad (28)$$

However, taking into account (20) and the Appendix of [6], we see that the second term of the right side of (28) is zero,

$$P\nabla_P(\tilde{S}(X,P)X^t) = 0. \quad (29)$$

From (28) and (29), we have

$$\nabla_P \tilde{g} = (\tilde{S}(X,P)X^t)^t$$

and (27) becomes

$$P(r+1) = P(r) + \begin{cases} 0, & \tilde{g}(r) > 0 \\ a_r(\tilde{S}(X(r),P(r))X(r)^t)^t, & \tilde{g}(r) \leqq 0 \end{cases} \quad (30)$$

If we use other criterion functions, different but similar algorithms are obtained. For example, if we let

$$R = \frac{1}{8}[|\tilde{g}(X,P)| - \tilde{g}(X,P)]^2, \quad (31)$$

then we have an algorithm called the relaxation algorithm,

$$P(r+1) = P(r) + \begin{cases} 0, & \tilde{g}(r) > 0 \\ a_r(\tilde{S}(X(r),P(r))X(r)^t|\tilde{g}(r)|, & \tilde{g}(r) \leqq 0. \end{cases} \quad (32)$$

Another example of the criterion function is

$$R = \frac{1}{2} \sum_{X \in C_1 U C_2} \{|\tilde{g}(X,P)| - \tilde{g}(X,P)\}. \quad (33)$$

The learning algorithm based on (33) is easily obtained by use of (33) in (25) but it is omitted.

VI. CONCLUSION

In order to improve the pattern classification power of the committee machine and to introduce an analytical expression of the general piecewise linear discriminant function, we have proposed the two-level committee machien. The two-level committee mahcine realizes the piecewise linear discriminant function in its general form. An analytical expression of the discriminant function of the two-level committee machine was presented, and a learning algorithm was given.

REFERNCES

[1] T.M. Cover et al., Nearest neighbor pattern classification, IEEE Trans., vol.IT-13 (1967).

[2] E.A. Patrick et al., A generalized k-nearest neighbor rule, Inform. Contr. vol.16 (1965).

[3] C.M. Ablow et al., A committee solution of the pattern recognition problem, IEEE Trans., vol.IT-11 (1965).

[4] W.C. Ridgway, An adaptive logic system with generalizing property, Stanford Eletron. Lab. Tech. Rep. 1556-1, Stanford Univ., (1962).

[5] M. Osborne, Seniority logic, IEEE Trans., vol. C-26 (1977).

[6] R. Takiyama, A general method for training the committee machine, Patt. Rec., vol.10 (1978).

[7] J.A. Cadzow, Synthesis of nonlinear decision boundaries by cascaded threshold gates, IEEE Trans., vol.C-17 (1968).

[8] O.L. Mangasarian, Multisurface method of pattern separation, IEEE Trans., vol.IT-14 (1968).

[9] J.L. Francalangia, A general class of layered threshold logic networks, Ph.D.Thesis, Univ. of Washington (1968).

[10] R.Takiyama, Multiple threshold perceptron, Patt. Rec., vol.10 (1978).

[11] R. Takiyama, A learning procedure for multisurface method of pattern separation, Patt. Rec. vol.12 (1980).

[12] C.L. Chang, Pattern Recognition by piecewise linear discriminant functions, IEEE Trans. vol.C-22 (1973).

[13] R. Takiyama, A two-level committee machine: A representation and a learning procedure for general piecewise linear discriminant functions, Patt. Rec., vol.13 (1981).

PROBABILISTIC CLUSTER LABELING OF IMAGERY DATA*

T. C. Minter, R. K. Lennington, and C. B. Chittineni

Lockheed Engineering and Management Services Company, Inc.
1830 NASA Road 1, Houston, Texas 77058

ABSTRACT

In this paper, the authors condiser the problem of obtaining the probabilities of class labels for the clusters using spectral and spatial information from a given set of labeled patterns and their neighbors. A relationship is developed between class and cluster conditional densities in terms of probabilities of class labels for the clusters. Fixed-point iteration schemes are developed for obtaining the optimal probabilities of class labels for the clusters. These schemes utilize spatial information and also the probabilities of label imperfections. Furthermore, experimental results from the processing of remotely sensed multispectral scanner imagery data are presented.

INTRODUCTION

Recently, considerable interest has been shown in developing techniques for the classification of imagery data such as remote sensing data obtained using the multispectral scanner (MSS) on board the Landsat for inventorying natural resources, monitoring crop conditions, detecting mineral and oil deposits, etc. Usually, the inherent classes in the data are multimodal, and nonsupervised classification or clustering techniques[1-3] have been found to be effective[4,5] in the classification of imagery data. Clustering the data partitions the image into its inherent modes or clusters. Labeling the clusters is one of the crucial problems in the application of clustering techniques for the classification of imagery data.

Cluster labeling is similar to the problem of labeling the regions obtained by using segmentation algorithms in the development of scene understanding systems. The recent literature shows considerable interest in the use of relaxation labeling algorithms for labeling the segmented regions.[6-8] These algorithms use relational properties of the regions through compatibility coefficients. In cluster labeling, the relational properties of the clusters are either not available or not meaningful. For example, in aerospace agricultural imagery, the regions of interest are crops, nonagricultural areas, etc. These can be anywhere in the image. Hence, it is not meaningful to define relational properties for the clusters.

Most of the imagery data contain much spatial information, and several researchers[9-12] have attempted to use spatial information in the classification of imagery data.

This paper documents an investigation of the problem of labeling the clusters using spectral and spatial information. It is assumed that the probability density functions and a priori probabilities of the clusters or modes are given. Let these respectively be $p(X|\Omega = i)$ and δ_i; $i = 1,2,\cdots,m$, where m is the number of modes or clusters. It is also assumed that a set of labeled patterns $X_i(j)$ with labels $\omega_i(j) = i$ and their neighboring patterns $Y_i^k(j)$ are given ($k = 1,2,\cdots,\ell$; $j = 1,2,\cdots,N_i$; and $i = 1,2,\cdots,C$, where C is the number of classes).

In remote sensing, the labels for the patterns are provided by an analyst-interpreter (AI), who examines imagery films and uses other data such as historic information and crop calendar models. Very often the AI labels are imperfect. Recently, Chittineni[13-15] investigated techniques for the estimation of probabilities of label imperfections using imperfectly labeled and unlabeled patterns. It is assumed that the probabilities of label imperfections are available. Methods are developed in the paper for obtaining probabilities of class labels for the clusters using all the available information.

A RELATIONSHIP BETWEEN CLUSTER AND CLASS CONDITIONAL DENSITIES

In this section, a relationship is developed between cluster and class conditional densities. In general, the class conditional density functions are multimodal. Let C be the number of classes and m be the number of clusters. Let $p(X|\omega = i)$ be the class conditional densities and $p(X|\Omega = i)$ be the mode or cluster conditional densities. Let $P(\omega = i)$ and $P(\Omega = i)$ be the a priori probability of class i and the a priori probability of cluster i, respectively. The mixture density $p(X)$ can be written in terms of class conditional densities as follows.

*The material for this paper was developed and prepared under Contract NAS 9-15800 for the Earth Research Division at the National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston, Texas.

$$p(X) = \sum_{i=1}^{C} P(\omega = i)p(X|\omega = i) \quad (1)$$

The mixture density p(X) can also be written in terms of mode conditional densities as

$$p(X) = \sum_{\ell=1}^{m} P(\Omega = \ell)p(X|\Omega = \ell)$$
$$= \sum_{\ell=1}^{m} p(X|\Omega = \ell) \sum_{i=1}^{C} P(\Omega = \ell, \omega = i)$$
$$= \sum_{i=1}^{C} P(\omega = i) \sum_{\ell=1}^{m} P(\Omega = \ell|\omega = i)p(X|\Omega = \ell) \quad (2)$$

The following assumption is made from comparing equations (1) and (2).

$$p(X|\omega = i) = \sum_{\ell=1}^{m} P(\Omega = \ell|\omega = i)p(X|\Omega = \ell) \quad (3)$$

Equation (3) can be rewritten as

$$p(\omega = i|X) = \sum_{\ell=1}^{m} \alpha_{\ell i}p(\Omega = \ell|X) \quad (4)$$

where $\alpha_{\ell i} = P(\omega = i|\Omega = \ell)$ and is the probability that the label of mode ℓ is class i. The probabilities $\alpha_{\ell i}$ satisfy the constraints given in equation (5).

$$\alpha_{\ell i} \geq 0 \;\; ; \;\; i = 1,2,\cdots,C \text{ and } \ell = 1,2,\cdots,m$$
$$\sum_{i=1}^{C} \alpha_{\ell i} = 1 \;\; ; \;\; \ell = 1,2,\cdots,m \quad (5)$$

Equation (3) provides a relationship between class and cluster conditional densities in terms of probabilities of class labels for the clusters.

MAXIMUM LIKELIHOOD PROBABILISTIC CLUSTER LABELING

This section concerns the problem of obtaining the probabilities $\alpha_{\ell i}$ (the probabilities of class labels for the clusters). It is assumed that we are given a set of labeled patterns $X_i(j)$ with class labels $\omega_i(j) = i$; $j = 1,2,\cdots,N_i$, and $i = 1,2,\cdots,C$. It is also assumed that the a priori probabilities of the modes or clusters and mode conditional densities are given. Let δ_i and $p(X|\Omega = i)$ be the mode a priori probabilities and mode conditional densities, respectively. The criterion used in obtaining the probabilistic description of class labels for the clusters is the likelihood function. The likelihood of an occurrence of patterns $X_i(j)$ with their labels $\omega_i(j) = i$ is given by

$$L_1' = \prod_{i=1}^{C} \prod_{j=1}^{N_i} p[X_i(j), \omega_i(j) = i] \quad (6)$$

Since $\prod_{i=1}^{C} \prod_{j=1}^{N_i} p[X_i(j)]$ is independent of $\omega_i(j)$, for mathematical simplicity, dividing the above equation by it yields

$$L_1 = \prod_{i=1}^{C} \prod_{j=1}^{N_i} \frac{p[X_i(j), \omega_i(j) = i]}{p[X_i(j)]} \quad (7)$$

Noting that the logarithm is a monotonic function of its argument and taking the logarithm of L_1 of equation (7) and using equation (4) yield the following.

$$L = \log(L_1) = \sum_{i=1}^{C} \sum_{j=1}^{N_i} \log\left\{\sum_{\ell=1}^{m} \alpha_{\ell i}p[\Omega = \ell|X_i(j)]\right\} \quad (8)$$

The probabilities $\alpha_{\ell i}$ satisfy the constraints given in equation (5). Closed-form solutions for $\alpha_{\ell i}$ by maximizing L of equation (8), subject to the constraints of equation (5), seem to be difficult to obtain. The probabilities $\alpha_{\ell i}$ can easily be obtained using optimization techniques such as the Davidon-Fletcher-Powell procedure.[16-18]

The following fixed-point iteration equation (similar to maximum likelihood equations in parametric clustering[3]) for the solution of the above optimization problem can easily be obtained by introducing Lagrangian multipliers. That is,

$$\alpha_{\ell i} = \frac{\sum_{j=1}^{N_i} d_{\ell ij}}{\sum_{i=1}^{C} \sum_{j=1}^{N_i} d_{\ell ij}} \quad (9)$$

where $$d_{\ell ij} = \frac{\alpha_{\ell i}p[\Omega = \ell|X_i(j)]}{\sum_{s=1}^{m} \alpha_{si}p[\Omega = s|X_i(j)]} \quad (10)$$

However, closed-form solutions for $\alpha_{\ell i}$ can be obtained with the criterion as the maximization of a lower bound on L, and they are given in the following discussion.

$$\alpha_{\ell i} = \frac{N_i e_{i\ell}}{\sum_{r=1}^{C} N_r e_{r\ell}} \quad (11)$$

where $$e_{i\ell} = \frac{1}{N_i} \sum_{j=1}^{N_i} p[\Omega = \ell|X_i(j)] \quad (12)$$

This solution simply states that the probability of the ith class label for a given cluster ℓ is the ratio of the sum of the a posteriori probabilities of cluster ℓ, given the labeled patterns from class i, to the sum over all classes of the sum of a posteriori probabilities of cluster ℓ, given the labeled patterns from each class. Having obtained $\alpha_{\ell i}$, q_i (the proportion of class i) can be estimated as follows.